DEVELOPMENT OF TEST ITEMS

**U.S. Department of Energy
Washington, D.C. 20585**

FOREWORD

This document provides information to DOE staff and contractors that can be used by training staffs (e.g., instructors, designers, developers and managers) and others for the development of test items. This document has been prepared on the basis of test item development methods used at various DOE nuclear facilities. This document, used in conjunction with *Design, Development, and Implementation of Examinations* (available in this Library), can be used as an aid to provide a framework for the systematic development of valid and reliable examinations. Detailed examples of short answer, multiple-choice, and essay test-item formats are included to aid in developing effective test items.

Users are not obligated to adopt any part of this document; rather, they can selectively use the information to establish or improve facility training programs as applicable. This document was based upon DOE Handbook *Guide to Good Practices for the Development of Test Items.*

Beneficial comments (recommendations, additions, and deletions) and any pertinent data that may be of use in improving this document should be addressed in the Support Forum Section of this Site.

TABLE OF CONTENTS

1. INTRODUCTION

While the methodology used in developing test items can vary significantly, to ensure quality examinations, test items should be developed systematically. Test design and development is discussed in the *Design, Development, and Implementation of Examinations* (available in this Library). This document is intended to be a supplement by providing more detailed guidance on the development of specific test items.

This document addresses the development of written examination test items primarily. However, many of the concepts also apply to oral examinations, both in the classroom and on the job. This document is intended to be used as guidance for the classroom and laboratory instructor or curriculum developer responsible for the construction of individual test items.

This document focuses on written test items, but includes information relative to open-reference (open book) examination test items, as well. These test items have been categorized as short-answer, multiple-choice, or essay. Each test item format is described, examples are provided, and a procedure for development is included. The appendices provide examples for writing test items, a test item development form, and examples of various test item formats.

2. GENERAL DISCUSSION

Test items consist of two components: the content (what is asked) and the format (the way it is asked). The quality of the test item depends as much on the way the test item is constructed as on the content. Important topics that are tested by ambiguous, awkward, or poorly specified test items cannot be considered valid. Therefore, the selection of valid topics and the careful construction of test items are equally important parts of a valid test development process.

The conversion of a learning objective into a test item is, in part, a creative process. Yet, there are certain procedures and guidelines that can help in writing a test item and in ensuring that the test item will measure the knowledge, skill, or ability that it is intended to measure.

Five steps are involved in developing test items:

- Test item format should be determined

- Number of test items to be developed should be determined

- Skill and knowledge test items should be written

- Test item content should be validated

- Test items should be incorporated into the test bank for future use.

Determination of test item format should be guided by the action verb of the learning objective. However, any format of test item can be used to test any cognitive level if an appropriate stem is developed. Skill verbs suggest performance testing while knowledge verbs suggest one or more of the written formats. For example, "start up" and "shut down" are skill verbs that suggest a performance test, whereas "recall," "identify," and "list" are verbs more appropriately addressed by the completion and short-answer formats. Verbs such as "discriminate" and "select" should be tested using a multiple-choice format. Learning objectives that require the trainee to classify or relate are most easily tested by using the matching format. Higher-order verbs such as "synthesize" and "evaluate" may be tested using an essay format. Appendix B gives a detailed listing of the action verbs versus the domains and levels within the domains.

After the test item format is determined, the number of test items to be developed should be established. Some of the factors that significantly affect the number of test items to be developed are the number and type of examinations on which the test items will be used, the ease with which test items can be created, and how frequently a test item will be reused. The initial number of test items selected for development should range from at least one to several test items per learning objective.

Test items should be written after the format is selected and the number of test items is determined. Test items should have answer keys written at this time.

Content validation is the process by which test items are determined to be consistent between the content being learned and what is being tested. Each test item should be reviewed by subject matter experts (SMEs). These validation reviews should be documented.

A test item bank should be developed and maintained current. The test items in the test bank may be used to construct entry-level tests, progress tests, or after-instruction tests. A tracking system should be developed to correlate test items to the corresponding learning objective.

The following chapters offer guidance on how to select, construct, and score different formats of written test items (e.g., short-answer, multiple-choice, matching, and essay). Traditionally, test items that require the trainee to supply an answer (e.g., short-answer, essay) have been considered "subjective"; test items requiring the trainee to select an answer (e.g., multiple-choice, matching) have been considered "objective." If graders require subject matter expertise to interpret the answers of test takers, the test item is considered subjective. If the examination can be scored without having to interpret the answer (e.g., machine scored), it is considered objective. An objective test item is defined here as one in which: (a) there is only one correct answer, and (b) all qualified graders would agree on the amount of credit allowed for any given trainee's answer.

Writing test items that are technically sound is a skillful and challenging task. Exam writers should balance technical concepts with an array of test item writing principles to create a single "good" test item. The test item should be one that reasonably assesses what a trainee ought to know. The test writing challenge is further complicated by the inherent ambiguities and interpretations of the written language. To make it perfectly clear in the mind of the trainee what the test item is asking for, test items should be as precise and focused as possible.

Finally, the overall written examinations can be highly streamlined if examiners adapt use of a separate answer sheet for the recording of trainee answers. The separate answer sheet affords several key benefits.

- Directions for recording answers for multiple-choice test items need only be given once (at the onset of the exam) rather than repeating "circle the answer below" for each test item

- Examiner scoring can be made easier and more rapid through either the use of an overlay key or an optical scanner

- Test item analysis can be accomplished more readily through either a single-sheet, visual profile of responses or through a computerized test item analysis via optical scanner.

In translating a learning objective into a test item, the following examples may help the examiner generate ideas for potential test items. These examples not only help in the development of test items, but they may also be useful when generating learning objectives.

1. What are the common misconceptions about _____?
2. Why is _____ important to satisfactory job performance?
3. In what sort of circumstances might it be important to understand _____?
4. What might the individual do who does not understand_____ ?
5. What might the consequences be of a lack of knowledge about _____ ?
6. How can the individual demonstrate this knowledge?

3. DEVELOPMENT OF WRITTEN TEST ITEMS

Guidance on the construction of each test item format is provided in the following chapters. However, there are basic principles that apply across all test item formats.

- Ensure that the concept is relevant to the ability to perform the job

- State the test item concisely

- Choose the higher cognitive level

- Make sure the test item matches the learning objective

- Omit unnecessarily difficult or irrelevant test items

- Limit the test item to only one concept or topic

- Avoid copying text directly from reference materials

- Avoid backwards logic test items

- Place the easier test items at the beginning of each section

- The test item should discriminate between those who have mastered the objective and those who have not.

Ensure that the concept being measured has a direct relationship to the ability to perform the job. The construction of the test item should clearly reflect the enabling objective. Word the test item so that it would be considered valid and reasonable to other SMEs using the same reference materials.

State the test item as concisely as possible, but provide all necessary information. The test item should be clear, grammatically correct, and free of clues to the correct answer. It should be written at a reading level appropriate for the trainee. Often the individuals who develop a test item assume that certain conditions are inherent in the question when, in fact, they are not. It is important to have others review your test items to ensure that all necessary information is included, and that all excess information is deleted. You should ask yourself: Will the trainees clearly know what they are expected to do? Do they have all the information they need to answer the test item? Does answering the test item depend on certain assumptions that must be stated?

When there is a choice between two cognitive levels, write your test item to reflect the higher level. Learning objectives and test items should be written to reflect the level of cognitive domain that is most appropriate. Examinations should consist of higher-level cognitive test items.

Make sure that the test item matches the learning objective. It is very easy to end up with a test item that tests a relatively trivial aspect of an important learning objective. When reviewing your draft test item, ask yourself whether it is likely that someone could answer the test item correctly and still not meet the objective or perform the tasks.

Omit test items that are irrelevant. When reviewing your draft test item, ask yourself: could someone do the job safely and effectively without being able to answer the test item? If so, is it because the content is inappropriate, the wording is unclear, or the level of understanding is too great?

Limit the test item to one concept or topic, unless a synthesis of concepts is being tested. Each individual test item should be reserved for testing one topic, and that topic, as well as the intent of the test item, should be clear to both examiner and trainee. There is a common misconception that testing for multiple topics in one test item is a time-efficient way to examine. Test items containing a variety of topics only serve to confuse the trainee about the purpose of the test item and, therefore, what is expected in terms of a correct response.

Avoid copying text directly from training or other reference material. Test items written in this way generally encourage rote memorization. Further, copying from reference material can cause confusion in test items because the material lifted often draws its meaning (and importance) from its surrounding context. Therefore, important assumptions or conditions stated elsewhere in the material are often omitted from the test item.

Avoid "backwards logic" test items -- those test items that ask what should be provided in the test item, and provide what should be required in the trainee's response. It is important to test topics in a way consistent with how the topic should be remembered and used. For example, consider the following test item:

> Test item:
>
> If it takes 12.5 cubic feet of concrete to build a square loading pad 6 inches thick, what is the length of one side of the pad?

This test item gives the test takers information they should be asked to calculate, while it requires them to provide information that would be supplied in an actual work situation. In constructing your test items, make sure that you include information that trainees would typically have or have access to, and require responses that reflect the decisions, or calculations, or other information they would typically have to supply.

Place the easier test items at the beginning of each section. These test items help trainees gain composure and confidence. However, this is not to say that extremely easy test items should be included in the exam for the sole sake of relieving trainee tension.

Finally, a test item should be worded so that it discriminates between those who have mastered the objective and those who have not. A well-written test item should parallel the objective that it is testing.

4. SHORT-ANSWER TEST ITEMS

Short-answer items require the trainee to write a response in contrast to selecting from among a set of alternative responses. A short-answer item may be presented in such a way as to give the trainee freedom to express ideas or it may restrict a response to a given content area or answer format. This test item format is useful in testing knowledge, comprehension, and simple application. This format is more objective than essay-format test items; however, it is not free from subjectivity in scoring. The "guess-factor" is reduced from that of a multiple-choice test item because the chance of successfully guessing the answer is less. Because of varying trainee responses, the person grading the test item needs a definitive answer key and an understanding of the subject matter.

There are three common styles of short-answer test items: completion, fill-in-the-blank, and list/diagram completion. In each case, the trainee must supply the correct answer to receive credit. Because the developer of a short-answer test item often has clear ideas about expected answers, it is essential that model answers, acceptable alternatives, and associated grading criteria be written at the same time as the test item.

Examples of short-answer test items are:

Completion:

How many protons and electrons does the element hydrogen contain? _____

Fill-in-the-blank:

In a pressurized water reactor, fuel pellets of _____ are clad in _____; the primary

coolant is _____.   .

List/diagram completion:

Complete the blanks in the following list, indicating thermocouple type and material:

TYPE	MATERIAL
K	Alumel-_____
	Iron+ Constantan-_____
T	Constantan-_____
	Chromel+ Constantan-_____
	Platinum-
R or S	Rhodium+_____

These formats are easy to construct and score. However, the disadvantages frequently outweigh the advantages, since spelling errors, grammatical errors, and legibility complicate the scoring process. Test graders cannot anticipate all possible responses and synonyms for the responses they expect, and therefore, often have to make subjective judgments on the correctness of trainees' responses.

Short-answer test items can be used to elicit various abilities. These include:

- Recalling terminology, facts, symbols, and classifications
- Applying rules and making interpretations of data and other information
- Solving scientific and mathematical problems
- Manipulating symbols and balancing mathematical and chemical equations.

5. SHORT-ANSWER TEST ITEM WRITING PROCEDURE

As with all test item formats, more consistent results are attainable if certain guidelines and rules are followed when the test items are developed. Keep the following in mind when constructing short-answer test items:

- Provide clear, explicit directions

- Avoid excessive wording or information

- Ensure the answer key matches the test item

- Avoid giving away the answer

- Avoid trick test items

- Use the question form if possible

- Word each test item in specific terms with clear meanings so the intended answer is the only one possible, and so the answer is a single word, brief phrase, or number

- Avoid copying statements word for word from texts or classroom materials

- State the precision, numerical units, or degree of accuracy expected of the answer

- Word the test item to avoid irrelevant clues or specific determiners

- Word each test item so the blank is toward the end of the sentence

- Omit important rather than trivial words from fill-the-blank test items

- Avoid "butchered" or "mutilated" sentences; use only one or two blanks in a complete sentence

- If more than one blank is used, keep the blanks equal in length.

The test item directions should provide clear, explicit directions for answering the test item so that the trainee understands what constitutes a fully correct response. Choose words carefully to ensure that the conditions and requirements of the test item are appropriately conveyed. Words such as "evaluate," "outline," and "explain," can invite lots of detail that is not necessarily relevant.

The test item developer should avoid excessive wording or information in the test item. In an attempt to make a test item operationally oriented and/or meaningful, there is a tendency to add more information than required for a correct response. Below are two examples which illustrate this point.

First Example:

The term 'undermoderated' is used to describe one of the major characteristics of a typical reactor core. The effect that an increase in power would have on the moderator-to-fuel ratio, is to_____the ratio and to _____K_{eff}.

The beginning sentence of the above test item adds nothing but extra reading time.

Second Example:

Two isotopes, one with a half-life of 10 days and another with a half-life of 5 days, sit on opposite trays of a set of scales. Assuming the isotopes have identical atomic weights, and there is one Curie of each isotope, which way will the scales tilt (which weighs more)?

The situation in the above test item bears little relationship to a situation that an operator would encounter on the job.

Make sure that the answer key response matches (and is limited to) the requirements posed in the test item. The test item below only asks for a yes or no response, yet the answer key indicates that an explanation of that one-word response is required.

Test item:

Can a certified operator who has not been at the facility for a period of six months immediately resume duties in a licensed position?

Answer:

No. In any case the facility management would have to be notified before the operator resumed duties, and only after certification by an authorized representative of the facility or by a demonstration by the individual that his/her understanding of facility operations and administration is satisfactory.

The test developer should avoid giving away part or all of the answer by the way the test item is worded. In the test item below half credit is allotted for answering the first part of the test item "yes." However, a test-wise trainee can realize that the answer has to be yes, or else the second part of the test item would have read something like "If so, how? If not, why not?"

Test item:

If the letdown line became obstructed, could boration of the plant be accomplished shortly after a reactor trip to put the plant in cold shutdown? (1 point) If so, how? (1 point)

Trick test items should be avoided. Trick test items can occur unintentionally when the answer key does not precisely match the test item. The test item below asks for how, not if, the termination criteria change.

Test item:

How does the SI termination criteria change following an SI reinitiation?

Answer:

It does not change.

Examples of short-answer test items are found in Appendix C -- Test-Item Examples.

6. MULTIPLE-CHOICE TEST ITEMS

Constructing a quality multiple-choice test item has two major advantages. First, the scoring of objective test items is considerably more reliable and less time consuming than scoring subjective test items. Second, since the test item requires less time to answer, more test items can be used to test trainee knowledge. This will provide better content coverage, which will also increase test validity.

There are two distinct parts of a multiple-choice test item, the *stem* and the *choices, alternatives, responses,* or *options*. The stem is the introductory statement, and its functions are to ask a question, define a task, or state a problem to be solved. The alternatives, responses, or options are the listed responses to the stem. The alternative part may be viewed in two classifications, the response that is the one correct or best answer, and the remaining incorrect responses (called *distractors*).

The multiple-choice format of test item can be a good selection to:

- To assess a trainee's ability to discriminate and make choices
- To comprehend concepts, principles, and generalizations
- To make judgments about and choices among various courses of action
- To infer and reason
- To interpret new data or information
- To apply information and knowledge.

The multiple-choice test item is not a good selection to use when the trainer needs to measure:

- The trainee's ability for recollection under minimum prompting
- The ability to explain and give examples
- The ability to produce and express unique or original ideas
- The ability to organize personal thoughts
- The ability to display thought processes or patterns of reasoning.

Multiple-choice test items require trainees to select the best answer from a series of choices. This series of choices can use calculations, descriptions, arguments, estimates, sketches, etc. They are useful for testing knowledge, comprehension, and application of knowledge and theory. Since the answer is provided, the subjective aspect of scoring is eliminated. Furthermore, scoring can be conducted by mechanical means. If the stem is properly developed, multiple-choice test items can effectively test higher-level, cognitive abilities, such as, analysis, synthesis, and evaluation. The latter abilities are difficult to measure directly through any type of written test item; however, the essay-format of test item is preferred for this application.

The multiple-choice test item is the best choice where there are numerous test items covering a category of topics, and the objectivity of scoring is the most important consideration. For

instance, an examination that determines the trainee's selection for the job position. The multiple-choice test item can be used to test a greater variety of learning objectives and does not require the trainee to write out and elaborate their answers, minimizing the opportunity for less knowledgeable trainees to *bluff* or *dress up* their answers.

Multiple-choice test items should not be used when:

- The test item calls for a numerical answer, as in computational problems
- Writing the answer does not take any longer than using an answer sheet or marking the answer to the multiple-choice test item.

The development of multiple-choice test items from learning objectives involves three steps:

- Create the stem of the test item by forming a question of an incomplete sentence that implies a question
- Write the correct answer to the stem in as few words as possible
- Write distractors that are plausible to trainees lacking the degree of knowledge you want the test item to assess.

The following table provides tips for use in developing the stem of a multiple-choice test item.

TABLE 1. Tips for developing the stem of a multiple-choice test item

To Do	To Avoid
1. If possible, write as a direct question.	1. Avoid extraneous, superfluous, and non-functioning words and phrases that are mere *window dressing*.
2. If an incomplete sentence is used, be sure that • It implies a direct question • The alternatives come at the end (rather than in the middle) of a sentence	2. Avoid phrasing the test item so that the personal opinion of the trainee is an option.
3. Control the wording so vocabulary and sentence structure are at a relatively low and non-technical level.	3. Avoid textbook wording.
4. In test items testing definitions, place the word or term in the stem and use definitions or descriptions as alternatives.	4. Avoid *cluing* or *linking* test items (i.e., having the correct answer to one test item clued or linked to the correctness of the answer of a previous test item).
	5. Avoid negatively worded test items.

6.1 Styles of Multiple-Choice Test Items

Multiple-choice test items are the most common form of structured response-type test items. By definition, multiple-choice includes alternate-choice (true/false, yes/no, right/wrong) and matching. Both types vary only in the number of alternatives available.

Matching test items can be used where developers want to assess ability to draw relationships between a similar set of ideas. While matching test items can be objectively scored, care should be taken to ensure that associations made in each column are similar in nature and kind (i.e., conditions and procedures, causes and results, actions and individuals).

The most common multiple-choice form consists of a stem and the choices. The choices include the correct answer and several distractors. This form is illustrated in the following example:

<u>Test item:</u>

The atoms of a chemical element all have the same:	STEM
(a) Atomic number (Correct choice)	C
	H
(b) Relative atomic mass (Distractor)	O
	I
(c) Number of neutrons in the nucleus (Distractor)	C
	E
(d) Number of nucleons in the nucleus (Distractor)	S

There are various ways of producing and selecting distractors. The test item developer can independently (or in discussion with colleagues) determine possible distractors. An effective method is to develop a simple test item of a fill-in-the-blank statement. For example: What do all the atoms of a chemical element have in common? Then, give this test item to a group of trainees. Possible distractors that will emerge from their responses are usually more effective than those devised by the test-item developer, particularly since the distractors also identify trainee learning errors.

The selection of distractors should be based on the required depth of trainee understanding. One set of distractors may be appropriate only when a general understanding of a subject is required, while a different set of distractors is needed when an in-depth understanding is required (see the Model A example below).

Developing a good multiple-choice test item, written at the application level of the cognitive domain, which also adheres to good test item construction guidelines (i.e., valid, operationally oriented, plausible distractors) presents a challenge to the developer. To assist and guide you in designing multiple choice test items, the following four models are presented:

<u>Model A:</u>

(a)   Correct answer

(b)   Incorrect answer

(c)   Incorrect answer

(d)   Incorrect answer.

Model A depicts the traditional multiple-choice design format. This model shows one correct single word/phrase answer followed by three incorrect single word/phrase options. The length of all options should be similar.

> Model B:
>
> (a)   Correct answer
>
> (b)   Plausible misconception
>
> (c)   Incorrect answer
>
> (d)   Incorrect answer.

Model B is a variation of Model A where a plausible misconception is used as an incorrect answer. Again the length of all options should be similar.

> Model C:
>
> (a)   Correct answer with correct condition (e.g., because, since, when, if, etc.)
>
> (b)   Correct answer with incorrect condition
>
> (c)   Incorrect answer with incorrect condition
>
> (d)   Incorrect answer with incorrect condition.

Model C depicts an acceptable multiple-choice design that uses conditions with answers (a condition in a setting, event, cause/effect that may make the answer correct or incorrect). Notice that Model C shows only one correct answer with its correct condition.

> Model D:
>
> (a)   Correct answer
>
> (b)   Incorrect answer
>
> (c)   Correct answer with incorrect condition
>
> (d)   Incorrect answer with incorrect condition.

Model D illustrates an acceptable model when it may not be possible to create all options in uniform length. This model shows paired lengths -- two long and two short options -- which avoids setting any single option apart (either too long or too short) from the remaining options.

The four models presented are basic. Other models may be developed and used in combination with one another.

6.2 Examples of Multiple-Choice Test Items

Four-Option Multiple-Choice Test Item -- A multiple-choice test item with the same correct answer (key) can be written in several ways depending on the breadth of knowledge being assessed. The following example shows three variations of a multiple-choice test item. (The correct answer in each is "rad," but the distractors have been chosen with different objectives in mind):

Variation 1:

In health physics, the unit of absorbed dose is:

(a) Volt   (b) Rad   (c) Calorie   (d) Newton

The above set of choices covers such a wide range that discrimination would be extremely easy for those with only a vague understanding of health physics.

Variation 2:

In health physics, the unit of absorbed dose is:

(a) Rad   (b) Rem   (c) Kerma   (d) Disintegration/second

The above set of choices covers such a narrow band of knowledge that only those who have a good understanding of health physics would choose the correct answer.

Variation 3:

In health physics, the unit of absorbed dose is:

(a) Counts/second   (b) Rad   (c) Half-life   (d) Rem

The above set of choices covers two bands of knowledge. Counts/second and half-life would be immediately dismissed by those with some understanding of health physics, leaving the choice between (b) and (d) (i.e., one choice from two, not four, distractors).

7.   MULTIPLE-CHOICE TEST ITEM WRITING PROCEDURE

Although multiple-choice test items are not as easy to construct as other forms, they are very versatile, can be used to test all levels and types of knowledge, and minimize the likelihood of the trainee obtaining the correct answer by guessing. The following guidelines should be considered as you construct multiple-choice test items:

- Use four-answer options

- Do not use "none of the above" or "all of the above"

- Do not present a collection of true-false statements as a multiple-choice test item

- Define the test item, task, or problem in the stem

- Avoid negative stems

- Provide sufficient counterbalance in the distractors

- Include common misconceptions as distractors

- Make answers plausible and avoid overlapping answers

- Ensure that the distractors grammatically follow the stem.

The four-distractor, multiple-choice test item is the most common and is the style preferred. The five-answer option contributes nothing to the test item but confusion, and any format with fewer than four distractors makes guessing correctly more probable.

"All of the above" test items provide inadvertent clues to the trainee. When the "all of the above" option is the correct response, the trainee need only recognize that two of the options are correct to answer the test item correctly. When the "all of the above" option is used as a distractor, the trainee needs only to be able to determine that one option is incorrect in order to eliminate this option. "None of the above" responses should not be used with "best answer" multiple-choice test items, since it may always be defensible as a response.

Each test item should be focused on one topic. A test item containing answer options related to many separate issues does not increase the efficiency of the test item. To the contrary, test items with multiple topics only confuse the trainee about the meaning and purpose of the test item.

The test item, task, or problem should be defined in the stem of the test item. Include as much information about the problem or situation in the stem as necessary, leaving only the solution, action, or effect for the answer options:

Poor:

At 50% power

(a)   The equilibrium xenon reactivity worth is approximately equal to the equilibrium xenon worth at 100% power.

15

(b)   The equilibrium xenon reactivity worth is approximately one-half the equilibrium xenon worth as 100% power.

(c)   The equilibrium xenon reactivity worth is approximately two-thirds the equilibrium xenon worth at 100% power.

(d)   The equilibrium xenon reactivity worth is approximately three-fourths the equilibrium xenon worth at 100% power.

<u>Better:</u>

Which of the following approximates the 50% power equilibrium xenon reactivity worth to the 100% power equilibrium xenon reactivity worth?

(a)   Equal to

(b)   One-half

(c)   Two-thirds

(d)   Three-fourths.

Negative stems should be avoided because they are unnecessarily difficult to interpret. Negatively stated stems can also lead to double negatives. It is more important to test what something is than what it is not. If a negative stem is necessary, highlight the negative word (e.g., not, ever, least).

It is very tempting to write negatively stated test items, since they can be constructed by picking three true statements out of the reference material and changing a fourth statement to make it false. However, studies have shown that trainees do not do as well on negatively stated test items, either because they overlook the negative word or because negatively stated test items require trainees to pick an answer or characteristic that is not true, which can be somewhat confusing. In addition, these test items tend to emphasize negative learning. For example, consider the following stem of a multiple-choice test item:

<u>Test item:</u>

During 100% power operation, the feedwater dump valve opens inadvertently. The condensate pumps will not do which of the following?

This stem can be made to read positively as follows:

<u>Test item:</u>

During 100% power operation, the feedwater dump valve opens inadvertently. The condensate pumps will:

(a)   Increase flow to maintain feedwater flow rate.

(b)   Trip due to a runout condition.

(c)   Have no response.

(d)   Trip due to low suction pressure.

There are times when a negatively stated test item is unavoidable. However, <u>never</u> use a negatively stated stem with a negatively stated answer option:

<u>Test item:</u>

Which of the following indications would <u>not</u> be expected and might indicate an instrument failure?

 (a) The CRD 'travel' lamp does not indicate when Group 8 rods are in motion.

 (b) Group 7 out-motion is prevented past 91.4%.

 (c) When you depress the 'CRD travel in' lamp-test push button, the 'CRD travel out' lamp comes on.

 (d) During a transfer of a group from direct current, hold to auxiliary; when you select 'SEQ-OR,' the 'SEQ-OR' lamp is on and the 'SEQ' lamp goes off.

Notice how confusing the option (a) answer is in combination with the test item stem.

Provide sufficient counterbalance in test items with multipart answers. Multiple-choice test items can legitimately contain multipart answer options. However, if the answers contain too many parts and/or too many options for each part, cues indicating the correct answer may be unavoidable. Consider the following example:

<u>Test item:</u>

The RCS is in hot standby with no reactor coolant pumps running. If OTSG pressure is decreased, according to the plant verification procedure, which of the following temperature responses indicate the presence of natural circulation?

 (a) T_h increases, T_c remains the same.

 (b) T_h increases, T_c decreases.

 (c) T_h decreases, T_c decreases.

 (d) T_h remains the same, T_c decreases.

The trainee could choose the correct answer (c) without knowing about the T_c temperature response in this situation, since "T_h decreases" only occurs in option (c).

Notice that two-part answers, with each part containing a two-option response, provides complete counterbalance, since all contingencies can be covered in four responses. For example:

<u>Test item:</u>

Which of the following is a definition of quadrant power tilt ratio?

 (a) Minimum upper detector output divided by average upper detector output.

 (b) Maximum upper detector output divided by average upper detector output.

 (c) Minimum upper detector output divided by average lower detector output.

 (d) Maximum upper detector output divided by average lower detector output.

A multipart test item which is highly recommended is one in which the two-part answer options consist of a two-level response (e.g., yes/no; off/on) and a reason. For example:

Test item:

Which of the following best describes the behavior of equilibrium xenon reactivity over core life?

(a) It decreases, because of the increased fuel burn up.

(b) It decreases, because of the decrease in plutonium-xenon yield.

(c) It increases, because of the increase in thermal flux.

(d) It increases, because of the decrease in boron concentration.

When possible, include common misconceptions as distractors. Since the purpose of the examination is to differentiate between competent and less-than-competent trainees, a good source of test items involves topics in which there are common misconceptions about important knowledge topics. For example, the following test item was based upon a common misconception about loss of subcooling margin:

Test item:

During a LOCA with a resultant loss of subcooling margin, why are the reactor coolant pumps (RCPs) secured?

(a) To prevent pump damage resulting from operation under two-phase conditions.

(b) To prevent core damage resulting from separation upon subsequent loss of RCS flow.

(c) To reduce RCS pressure by removing the pressure heat developed by the RCPs.

(d) To remove the heat being added to the RCS by the operating RCPs.

Make all answer options homogeneous and plausible. For example, the choices of answers to the following test item illustrates these characteristics:

Test item:

On a loss of condenser circulating water intake, the upper surge tank, hotwell, and condensate storage tank will supply sufficient feedwater to allow decay heat removal for approximately:

Poor	Better
15 minutes	8 hours
8 hours	24 hours
48 hours	48 hours
3 months	72 hours

One method of changing the difficulty level of a test item is to vary the similarity among answer options. Develop distractors that are similar enough to be chosen by those who do not meet the

testing objective, yet different enough so that they do not test trivial issues or distinctions. If the answer options have a logical sequence, put them in order.

Avoid overlapping answer options:

<u>Test item:</u>

The SPND uses rhodium which decays with a half-life of 42 seconds. How long will it take for a detector to indicate approximately 95% of an instantaneous power level change?

<u>Poor</u>	<u>Better</u>
2-4 minutes	1-2 minutes
4-6 minutes	3-4 minutes
6-8 minutes	5-6 minutes
8-10 minutes	7-8 minutes

To prevent giving the correct answer, ensure that the distractors do follow grammatically from the stem -- for example:

<u>Test item:</u>

During 100% normal power operation a single steam flow element to the steam generator feedwater fails high. The steam generator feedwater control system will cause:

  (a)   The feedwater valves to increase steam generation level slightly before returning the level to normal.

  (b)   Before returning the level to slightly above normal, the feedwater valves to increase the steam generator significantly.

  (c)   The feedwater valves to increase the steam generator level to the level of a reactor trip.

  (d)   The feedwater valves to increase the steam generator level slightly and maintain the increased level.

<u>Test item with Improved Distractors:</u>

During 100% normal power operation a single steam flow element to the steam generator feedwater fails high. The steam generator feedwater control system will cause:

  (a)   The feedwater valves to increase steam generator level slightly before returning the level to normal.

  (b)   The feedwater valves to increase the steam generator level significantly before returning the level to slightly above normal.

  (c)   The feedwater valves to increase the steam generator level to the level of a reactor trip.

(d)   The feedwater valves to increase the steam generator level slightly and maintain the increased level.

8. MATCHING TEST ITEM WRITING PROCEDURE

With matching test items, a trainee is required to match each word, sentence, or phrase in one column with a word, sentence, or phrase in another column. The test items in the first column are called "premises." The answers in the second column are called "responses." Consider the following guidelines when constructing matching test items.

- Give specific directions for each matching test item that indicate on exactly what basis the matching is to be done.

- State in the directions whether a response can be used only once or more than once.

- Only use logically related material in a single matching test item -- nothing should be listed in cross-related columns that is not part of the subject of the test item.

- Consider the following possibilities for pairing premises and responses:

PREMISES	RESPONSES
Terms or words	Definitions
Short test items	Answers
Symbols	Proper names
Causes	Effects
Principles	Situations in which they apply

- For any single matching test item, use only one set of premise-response formats (e.g., only terms with definitions).

- Each response should be a plausible answer for each premise -- for example:

Test item:

Match the alarm (letter) with its appropriate tone (number). Responses (numbers) may be used more than once:

PREMISES	RESPONSES	
A. Fire	1. Pulse Tone	_____
B. Reactor Building Evacuation	2. Wailing Siren	_____
C. Site Evacuation	3. Steady Tone	_____

- Present an unequal number of premises and responses, or allow responses to be used more than once -- this will inhibit the trainee from obtaining clues to a correct answer through a process of elimination.

- Arrange the responses in a logical order, such as alphabetical or numerical.

- The entire matching test item should contain no more than six premises.

- Place the entire matching test item on one page so the trainee need not flip back and forth between pages.

Directions to matching test items should <u>not</u> require more than one response to a premise. This practice not only fragments scoring and alters the implicit equality of the premises, but also may encourage "shotgunning," the writing in of all responses to a premise in order to receive some partial credit.

If partial credit is possible, establish how it will be assigned in the examination key. For example, a five-premise matching test item might have 20% credit for each correct match.

9. ESSAY-FORMAT TEST ITEMS

Test items that are called essay-format include those requiring the trainee to provide calculations, graphs, descriptions, arguments, estimates, sketches, or combinations of these. The trainee has to determine the answer and select the means of communication. Essay-format test items should be used when the need is to have the trainee express himself/herself or to recall information from memory rather than to present the information to the trainee for recognition or identification. Test items of this format can test the whole range of abilities, but they usually are considered most suitable for application, analysis, synthesis, and evaluation. The test item and the associated grading criteria should be prepared concurrently. The examiner grading the test item should be knowledgeable in the subject matter.

9.1 Styles of Essay-Format Test Items

There are two major styles of essay-format test items: structured and unstructured. In each case, trainees must supply the correct answer to receive credit. However, the unstructured test items require a greater amount of analysis and composition than are required by structured test items.

The unstructured essay-format test item is appropriate where a learning objective expects a trainee to organize ideas, to develop a logical argument, to present evaluations of certain thoughts, to communicate thoughts and feelings, or to demonstrate other abilities requiring original written expression. An example of this format would be an assignment for the trainee to "write a paper explaining the pro's and con's of using written procedures." A disadvantage of the unstructured essay-format test item is that scoring the response in an objective manner is difficult and time-consuming.

The structured essay-format test item does not have to be limited to testing for recall and comprehension. It can be designed to require the trainees to apply their knowledge and skills to solve new problems or analyze a novel situation. One way to do this is to write test item statements that require the trainee to apply their knowledge to specific material given in the test. For instance, a description of a particular, along with the initial facility conditions, can be given in the test item statement. One advantage of the structured essay-format test item is that by narrowing the focus of the test item statement to elicit specific and well-defined behaviors, the trainees are more likely to interpret the statement as the test developer intended. The structured essay-format test item also prompts the test developer to provide more clarity in presenting the correct answers, which can lead to improved reliability of scoring.

A disadvantage of all essay-format test items lies in their application, since they are often easier to develop than the multiple-choice test item. A good application of the essay-format test item is to offer the trainee an opportunity to exhibit abilities for written production, organization, expression, and interrelationships among ideas. A poor use of the essay-format test item is to test for factual information. The trainer can make a more appropriate test item format selection by analyzing whether the related objectives require the trainee to *recall and supply* answers or to *recognize* already supplied answers. If the latter applies, the multiple-choice test items are preferable because they allow for more objective scoring.

9.2 Examples of Essay-Format Test Items

It is possible to express the same test item as either a structured or unstructured test item. Two extremes are shown here:

<u>Structured Essay-Format Test items:</u>

A pressurized water reactor at full power has the following conditions:

      (a)   Core flow (100%)

      (b) T_{avg} (588 ° F)

      (c)   Core AT (60'F)

      (d)   Steam Generator AT $_{(T_{avg}-T_{steam})}$ (44°F).

The reactor is operating at steady-state power with four loops operating. One reactor coolant pump trips, resulting in a three-loop operation. The rod control system is in automatic. The reactor has not tripped.

      (a)   Sketch graphs of RCS mass flow, reactor power, core OT, and steam generator pressure versus time if no operator action is taken. Use the same time axis for all graphs. (4 points)

      (b)   Calculate the final reactor power, $T_{avg,}$ core OT, and steam generator OT. (4 points)

This test item gives the trainee clear guidelines on what is expected. This is termed a structured test item and requires little personal initiative, but a clear requirement exists to apply knowledge gained from theoretical and practical experience in the course. Also, this test item lends itself to more consistent grading, and is, therefore, more objective.

<u>Unstructured Essay-Format Test item:</u>

A pressurized water reactor is operating at 100% steady-state power. One reactor coolant pump trips resulting in a three-loop operation. Explain what happens if the reactor does not trip and no operator action is taken. State any assumptions you make. (8 points)

This test item requires the trainee to interpret the word "explain." They need to decide what the important issues are, what assumptions to make (possibly based on parameters taken from their own experience) and then communicate the issues and assumptions by whatever means they think are appropriate. Some trainees may use graphs and a short explanation; others may use only an explanation. Each answer will be unique. This test item requires the trainee to clearly express understanding of the topic.

Unstructured essay-formats are more difficult to grade consistently because of the variations that are possible, relative to assumptions the trainee makes. Consequently, the grading becomes more subjective in test items such as the one in the foregoing example.

For assessment purposes, most test-item developers prefer to write structured essay-format test items. Unstructured essay-format test items are most appropriate in learning exercises in

which feedback and additional clarification can be provided if necessary (oral examination, on-the-job training, walk-through, etc.).

10. ESSAY-FORMAT TEST ITEM WRITING PROCEDURE

The layout of the test item and the clarity of the wording are crucial for the trainee to understand what is required when essay-format questions are used. Generally the words "how" and "why" are the best, followed by "what" and "when," although the latter two can more easily lead to a single word or very short answers.

Trainees find particular difficulty with interpretation of words such as "discuss," "compare," and "explain." The level of the answer expected is important in determining whether a more specific statement is useful.

The test item developer should follow a systematic method such as the following:

- Define the behavior the trainee is expected to exhibit or describe the process to be exhibited before beginning to write the essay-format test item.

- Allow adequate time for developing the test item, model answer, and grading criteria. Essay-format test items usually require several revisions during the development process. Consider stating the anticipated answering time for the test item.

- Referring to the learning objectives, select a topic from which the test item is structured. Ask only for important relevant information.

- Ensure the test item is within the educational maturity level of the trainees.

- Determine the level of learning expected in response to a test item on this topic from the learning objective. Use the words "discuss," "explain," or "compare" only when deliberately aiming at higher cognitive levels.

- Define the problem explicitly -- an essay-format test item is of no use if the trainee cannot comprehend the test item. It should be open to only one interpretation so all trainees address the same topic.

- Ask test items that require the trainee to demonstrate the ability to use essential knowledge and do so in situations that are new or novel for the trainee, rather than simply recalling information.

- Limit the problem -- the scope of a content area should not be too large. Remember that trainees will take longer to answer the test item than the developer.

- Word the test item so the trainees know the limits of the tasks, their purposes, and can answer them in the time allotted.

- Ask questions that are relatively specific or focused, and which require relatively brief responses.

- If additional guidance for a particular test item is necessary beyond that shown in the general instruction, make a clear and concise statement.

- The subject matter experts should also agree on the relevance or importance of a particular test item to course completion, as well as the correct response.

- Develop the test item grading criteria. Identify the key responses expected from the trainee. Identify each necessary response (e.g., by underlining key words, phrases, or steps, or by circling important parts of graphs or diagrams). Determine the percentage value to be allocated for each test item subpart (based on their relative importance).

- Check the test item against the grading criteria:

  - Does the test item clearly request what is considered important in the grading criteria?

  - Is the time necessary to answer the test item proportional to the importance of the test time? (Critical objectives justify more time if required.)

  - Do trivial details (e.g., an extended calculation) take too much time to address?

  - How could the test item and/or the grading criteria be revised to be more consistent?

- Ensure that the sequence of requirements is in a logical order.

- Ensure that all the data (stimulus material) needed by the trainee is provided and clearly displayed together.

- Avoid test items that can be answered by a single word or short phrase.

- Avoid negative statements in test items; they may lead to double negatives that are difficult to interpret.

- State clearly whether exact answers or estimates (precise figures or sketches) are required.

- Ensure that the point value for each test item or part of a test item is stated.

- Write the test item on a form such as that shown in Appendix A -- Test Item Development Form.

- Word the test item so that trainee can judge the approximate length of the answer desired and knows the point values or weight each will be given.

- Have another test item developer review the test item for accuracy, format, and readability.

- Word the test item so the SMEs can agree on the correctness of a trainee's response.

11. TEST ITEM ANSWERS AND POINT VALUE

For each test item, a correct answer should be supplied in order to develop an answer key. The answers should be unambiguous so that the grading of the test item will be consistent. The minimum response for full credit should be given, and indications of the relative value of partial responses should be made. This kind of information is obvious in multiple-choice and matching test items, but short-answer and essay test items do leave room for interpretation.

To prevent interpretive inconsistencies in grading short-answer or essay test items, the designated instructor should include the minimum answer required for full credit. If the exact wording or specific concept should be included to fulfill the requirement of a correct answer, one method that may be used is to underline the necessary parts and assign point value to the individual parts. If the exact underlined wording is not necessary, include a note to the grader that the content is necessary, but the wording is left to the grader discretion. For example:

> Test item:
>
> State the purpose of the emergency diesel generator. (4 points)
>
> Answer:
>
> Provides emergency power (2.0 points) to essential loads (1.0 point) in the event of a total loss of power (1.0 point). NOTE: *Alternate wording acceptable—grader discretion requested.*

After the test item and answer key are written, the point value should be assigned in a clear and consistent manner. The assignment should be based on a point scale with the "easiest" test item assigned lower point value and the "more difficult" test item assigned higher point value. The point value should be placed in parentheses at the end of the test item stem.

11.1 Cognitive Level vs. Point Values

To determine ease or difficulty of a test item, consider the criticality of the enabling objective that is being tested to the overall mastery of the terminal objective which will allow successful performance of the job task. Also consider the cognitive level at which the objective being tested is written. For example, if you are writing multiple-choice test items for three enabling objectives, two of which are written at the knowledge level of Bloom's Taxonomy to support a third which is written at the comprehension level, then the point value assigned to the higher cognitive level would be higher than the point value assigned to the lower cognitive level. For example, this is how these concepts might be applied:

ACTION VERB	COGNITIVE LEVEL	POINTS
E.O. 1.1 List	Knowledge	(2.0 pt)
E.O. 1.2 State	Knowledge	(2.0 pt)
E.O. 1.3 Determine	Comprehension	(4.0 pt)

If multiple-choice test items are written for these action verbs, the first two test items would have a lower point values than the third. This is because the third is more "important," more critical to the mastery of the terminal objective (the first two objectives support the third). Also, the third

action verb is at a higher cognitive level (comprehension) than the cognitive level of the first two (knowledge).

Other considerations for determining point value of test items include the following:

- Impact on successful performance of the job task if the answer is not known by the trainee

- Learning level required (lowest level being knowledge and the highest level being evaluation)

- Number of answers required for the test item

- Relative degree of difficulty.

Another consideration in the assignment of point value is related to the test item. Test items having multiple parts requiring answers should have at least one point per answer, and matching test items should have at least one point per match. Point value allocations should be made and approved when drafted test items are submitted for review and approval.

11.2 Scoring Essay-Format Test Items

A major difficulty in the use of essay-format test items is the inability to maintain consistent scoring. The answer should clearly identify the key points of each test item, and the grading criteria should be followed closely to ensure consistency of scoring. Effective grading criteria eliminate bias toward trainees (whose answers appear plausible, have neat handwriting, or are known to the instructor) by removing subjectivity and promoting objectivity.

During scoring, it is easy to drift from the established grading criteria. It is, therefore, necessary to check each answer against the answer and grading criteria for every test. It is also advisable for the same person to score one test item across all tests before scoring the next test item. This increases the consistency of the scoring process.

Also during scoring, the examiner may find that a large percentage of trainees incorrectly answer a test item. The examiner should then reevaluate the test item to determine if the test item could have been interpreted differently than what was intended or if there may have been other potentially correct answers.

12. USE OF HANDOUTS

The way in which handouts are incorporated into the exam has a bearing on the reliability of the examination. If a trainee has to search for the applicable drawing or curve, it adds to their frustration level and the time consumed in test administration. The format of the exam itself also contributes to the exam's reliability. Not only does it take more time to complete an examination that incorporates a multitude of administrative requirements, but also, it detracts from the trainee's ability to focus on just the exam test items.

Include handouts such as matrices or incomplete system drawings on which the trainee should transcribe his/her answer directly behind the page of the applicable test item. This will help ensure the integrity of the trainee's package of answers.

Include all other handouts, such as reference material, Technical Safety Requirement excerpts, etc., within a separate handout package. This package should be page checked with the trainees before exam administration. Booklets, rather than copied pages, should be used whenever possible.

Use of single-page/single-test item layout should be considered. This has proven to facilitate both the exam grading and the test taking processes. It also minimizes the administrative burden on the trainee, greatly reducing the chances of any test items being missed.

13. OPEN-REFERENCE TEST ITEMS

Training environments use several teaching aids that promote trainee learning. Many of these aids (e.g., textbooks, manuals, and student guides) assist the trainee in mastering the required subject matter during the initial training phase. The trainee is expected to comprehend the material and commit important facts to memory. However, there are many references (e.g., tables, charts, schematics, and procedures) trainees do not need to commit to memory (and should not) but have to interpret while performing their job.

To test trainee competence in the use of this material, test items that require the use of references to solve a problem or to reach a conclusion should be developed. This requires that the reference, or a sufficient subset of the reference, be provided to the trainee during administration of the test. This use of reference materials during testing is referred to as open-reference or open book testing.

The test developer should determine which references and their applications are necessary after reviewing the learning objectives and the test specifications. While the open reference test is essentially no different than other written tests, there are several points to consider when using this method.

- References are considered tools trainees use to solve a problem or reach a conclusion. The test items should not directly test knowledge of the references (this should be done in a closed book section); rather, it should test for proper use and application of the references.

- Open-reference test items should not be a "direct lookup" in a reference. For example, asking the specific setpoint for a protective action or the horsepower of a motor are direct lookups. Rather the reference material should be an aid in going through a series of knowledge-based decisions in order to arrive at a logical conclusion or answer.

- Open-reference test items should test the trainee's ability to locate, use, and apply the information found in the references.

- Test items should contain unique or varied circumstances that the trainees have not previously encountered. This makes the test item a true indicator of the trainee's ability to apply knowledge through the use of the references versus merely remembering an application from an earlier training session. Familiarity with routine applications of a reference can lower the learning level to simple recall.

- If a test item has references provided, it is not appropriate to make the test item more or less difficult than a comparable closed book test item. While very difficult test items may be useful in differentiating among the most able trainees, they are not appropriate for job or task qualification. Do not make an open reference test item easier or more difficult than other formats.

- Keep test item requirements as close to real life situations as possible. It is important the test items address the use of references in a context similar to that found in the

job environment. While providing a chart from a handbook is good, giving the trainees handbooks and requiring that the chart be found can be even better. Keep the content of open reference test items valid.

- Care should be taken in the development of open-reference test items. Under certain circumstances, these may become double jeopardy test items. In other words, if the test developer is testing the ability to use a graph vice locate it, the trainees will not be able to demonstrate use of the graph if they cannot find it.

An open-reference test item is the most challenging kind of test item to develop because it should challenge the highest cognitive level of the trainee. These test items are the result of a great deal of effort on the part of the developer. Several revisions and reviews by others are usually required to refine the test item into a quality test item that will challenge the trainee's decision making ability.